Memory Effects and Experimental Design

Tamara Miner Haygood &
Karla K. Evans

Introduction

Depending on the research question being asked and the means by which it is addressed, an observer's memory for a previously-encountered condition may be either a vitally important part of the study or a potential source of damaging bias. Therefore, depending on the nature of the experiment, different approaches to the issue of observer memory would be appropriate. This chapter will examine different types of experimental designs and discuss the role that observer memory may play in the outcome of the experiment, and how the design of the experiment may influence the role of memory. We write with the assumption that the relevant type of memory is memory for visual objects or scenes, but much of what is discussed is applicable to stimuli encountered in other ways, for example an experiment seeking to determine what kind of siren attracts the most attention to an approaching emergency vehicle.

Background

Human memory is a series of information processing systems working at different timescales, with different storage capacities, types of conscious access, active upkeep and mechanisms of operation (Baddeley, 2007). Memory allows us to store and retrieve information over time. There are different taxonomies of memory, but a well-accepted taxonomy that allows researchers to generalize regarding its content (e.g., visual, verbal or abstract information) divides memory into three types of memory systems: sensory memory, short-term memory (STM or working memory) and long-term memory (LTM) (Schacter, 1994).

The fastest decaying memory system is sensory memory, which stores sensory information for less than a second after it has been perceived and is an automatic response with limited capacity. Rehearsal or practice will not prolong sensory memory (Sperling, 1963). This memory system is also often referred to as iconic or echoic memory depending on the sensory information that it stores (visual or auditory).

STM, often referred to as working memory, allows for maintaining information without rehearsal for several seconds. It also has severely limited capacity regardless of whether information is visually or verbally represented (Baddeley, 1986). This system is believed to have the function of actively maintaining information while it is being manipulated to perform different cognitive tasks allowing us to engage in everyday thought activities (Baddeley, 2000). This system is most often studied using a change detection paradigm in which observers are presented with a various sets of visual stimuli with arbitrary combinations of features (e.g., 4 circles each with a different color occupying one part of the visual display) for a very brief period of time. After a brief delay, they see another display with the same elements, but one of the features of one element might

have changed (e.g., one of the circles has changed color or position), and they are asked to report if there was a change in a specified element or not. The ability to detect correctly if there was or was not change gives a measure of storage capacity since observers are required to maintain otherwise unrelated information in memory for a brief period of time and then compare that information to a new incoming set of information. This capacity to maintain and manipulate information has been shown to correlate with differences in fluid intelligence, academic achievement, and reading comprehension (Alloway, 2010; Daneman, 1980; Fukuda, 2010; Kane, 2001).

The capacity of STM or working memory can be increased through the process of chunking, or dividing information into meaningful groups, and that capacity can be maintained through rehearsal. For example, rehearsing or repeating over and over either audibly or inaudibly a list of items to buy will allow you to remember them until you are able to write down your shopping list.

Unlike STM, LTM is believed to be more of a passive store with large capacity and potentially unlimited duration (Tulving, 2000) maintained by stable and permanent changes in the neuronal connections across the brain. The long-term store is highly structured with multiple levels of representation ranging from individual items to concepts (Brady, 2011) and is characterized by two broad types of stores. One, referred to as semantic memory, maintains general world knowledge (e.g., facts, ideas, meanings, concepts) accumulated throughout our lives and pre-existing different modality representations that underlie our ability to perceive and recognize sensory input. For example, by the time a person in Western society is only a few years old, he or she will have encountered enough different chairs that a new chair will be recognized as a chair even though it may differ in style or color from all of the previously-encountered chairs. The other type of LTM store, episodic memory, is composed of stored autobiographical information or events that give a person the ability to recall and identify learned items based on previous encounters. These memories are tied to their contexts, times, places and associated emotions and depend on an awareness of self (Schacter, 1994). Episodic memory can also be viewed as a map that ties together information in semantic memory and a process that allows us to time travel. This is the memory that allows us to encounter a dusty beach chair in the garage and relive for a moment a golden afternoon spent lolling around on the beach with family during a summer vacation that may have happened years or decades earlier. It is also the memory that allows us to distinguish between categories of objects and specific objects. For example, when we see a black office chair, semantic memory or stored knowledge about the visual form and features of a chair in general allows us to recognize it as a chair, but if we see a chair again hours later, episodic memory enables us to decide whether this is the exact same chair we saw before or a different chair.

It is episodic LTM, specifically visual episodic memory, that we will concentrate on in this chapter. Episodic memory is viewed as a late-developing and early-deteriorating memory system, susceptible to neuronal dysfunction and possibly unique to (Tulving, 2002). The ability to store and recollect information and events that one encountered develops in late childhood and is one of the first abilities to decline as one ages. Visual episodic memory, sometimes referred to as visual recognition memory, denotes our ability to explicitly and consciously remember (re-experience) an image that we have seen before but that has not been constantly held actively in mind (Brady, 2011). There is no one single memory task that solely measures one memory

system, however, the remember-know paradigm (Gardiner, 2001) is most often used to assess the capacity of visual recognition memory. Assessing the capacity of visual recognition memory allows researchers to understand processing states of that system such as encoding and retrieval. The observers in this paradigm are initially asked to memorize a large number of images and are then asked to determine if a subset of images subsequently shown are the ones that they had seen before or are novel. The results of the retrieval response of the observers for this paradigm are based on observers remembering (having a specific recollection of an item or event encountered before) and knowing (familiarity with the item but lacking detailed contextual information) the visual material.

Based on studies using this paradigm, visual recognition memory is thought to have a very big capacity (Shepard, 1967) with the most notable study by Standing showing that observers demonstrated 98% accuracy in retrieval after seeing 10,000 real world images for 5 seconds each and 83% accuracy after several days (Standing, 1973). More recent studies have argued that the retrieval is with high fidelity (Konkle, 2010; Brady, 2008; Hollingworth, 2005). For example, Brady et al. (2008) report that, after memorizing 2500 objects over the course of 5 hours, observers were able to select the studied object with great accuracy from an old/new pair of images. They tested this ability with differing degrees of resemblance between the old and new object. Observers' accuracy was 92% when the new object was in a different category (e.g. old object was a giraffe and new object was a dining room table and chairs.) Their accuracy was 88% when the old and new objects were from the same basic category (e.g., two different dining room tables), and accuracy was 87% when the old and new objects were the same object in a different state (e.g. the same dining room table with all the chairs pushed in versus one chair pulled out [Brady, 2008].)

However, this does not mean that visual recognition capacity is the same for all visual content types and unaffected by attention during the encoding and retrieval process (Henderson, 2003). For example, increasing attentional load by asking observers to perform another concurrent cognitive task during encoding (e.g., counting back by threes from a set three-digit number while memorizing an image) reduces the number and the fidelity with which visual information is stored in long-term memory (Evans, unpublished). Reduced visual richness (e.g., diversity of features) and reduced distinctiveness of visual information also shrink visual recognition memory capacity. Images of open country, deserts, and waterscapes tend to be somewhat featureless and monotonous. They show a memorability score (i.e., percentage of correct recognitions by an observer) of only 61% as opposed to images with distinct visual features, like a human face, that averages 81% memorability score (Isola, 2014). Vogt and Magnussen showed that once distinctive details (e.g., a decorative door handle on an image of a door) were removed from the same picture, the observers' performance dropped from 85 to 65% correct recognition score (Vogt, 2007). Further, reducing semantic diversity of visual material by having all the material come from just one semantic category (e.g., doors) also significantly limits mnemonic capacity (Baddeley, 2017). For example, memory scores for recognition of images from one semantic category (e.g., forest scenes) reached only 67% as opposed to a score of 85% for the same number of images originating from four different sematic categories (forest, mountain, cityscape, beach) (Evans and Marom, 2016). Lastly, a related capacity modulating factor to semantic diversity is the level of abstraction of the visual stimulus or rather the lack of availability of stored knowledge about the stimulus. Observers shown two-tone abstract face images (Mooney faces) that they perceived as faces showed markedly better recognition relative

to the ones they did not perceive as faces but rather as mere splotches (Wiseman, 1974; Koutstaal, 2003) (Fig 1). The availability of both stored knowledge and a label to connect to a perceived visual stimulus provide an enhanced and structured encoding scheme that can increase mnemonic long-term capacity.

Figure 1. A is a photograph of author Karla Evans. It is easily recognizable as a picture of a human face. B is a stylized, abstract rendering of the photograph of Dr. Evans. C is an abstract figure not based on a human face. Although B is clearly not an actual photograph of a human being, it bears enough resemblance to a face that we believe most people would recognize it as human-like and would have an easier time remembering and recognizing it than C.

The scheme or model for encoding incoming visual representations also constantly changes based on our experiences and learning, and this constant revision of models forms the basis of expertise. More experience with an image class leads to a more sophisticated encoding model for visual material from that category. In support of this idea, all observers are quick to label or recognize objects at the basic level (e.g., it's a car), while experts are fastest labeling at the subordinate category levels (e.g., it's a 1955 Ford Thunderbird hard-top convertible) (Mervis, 1981; Jolicoeur, 1985). Therefore, expertise with an image class should also lead to increased memory capacity for the items from that class (Chase, 1982). There is ample evidence of this across different areas of expertise. It has been demonstrated that chess masters outperform amateur chess players in memory for chess configurations (Chase, 1973). This is true for other sports too, with professional athletes and experts in baseball and basketball exhibiting a significantly better memory for plays and other meaningful material from their domain of expertise compared to novices (Allard, 1980; Frey, 1976; Voss, 1980). Art historians and medical image interpreters (radiologists and cytologists) show the same enhanced ability for images from their domain of expertise (Vogt, 2007; Evans, 2011). This is true for long-term mnemonic ability in other modalities like music, where considerable musical training results in increased mnemonic capacity for musical and non-musical auditory stimuli (Cohen, 2011).

To summarize, the memory relevant to recognition of images or other stimuli that might be of interest in observer-performance studies is episodic LTM. This LTM encompasses retention of episodic information for any time period greater than a few seconds. Both remembering a field-trip your class took in 3rd grade and remembering what you ate for breakfast this morning depend on episodic LTM. This bank of memories has a very large storage capacity, and the capacity is even greater for stimuli related to one's field of expertise. The more one knows about a subject, the more detail one can perceive and process about an encountered object related to that subject, and the more likely one is to remember the object.

Experimental Designs that Depend on Memory

Three commonly used experimental designs that rely heavily on the ability of the observer to remember a previously encountered condition are alternative forced choice (AFC) experiments, rank order experiments, and sequential viewing experiments.

Anyone who wears corrective lenses has probably participated more than once in an AFC situation somewhat analogous to the methods used in an AFC experiment. One sits in the darkened office of an eye doctor, an ominous, large, black mask-like contraption covering one's face, while one stares out of just one of the mask's eye holes attempting to focus on the letters on an eye chart. The doctor plops a lens into the eye hole. Look at the letters in the 6th line down… are they clearer now…….. or NOW, as a new lens plops into the eye hole. One must choose which lens made the "e" sharper and clearer, and to do that, one must remember what that "e" looked like with the first lens, even when one is currently staring through the second lens. Once one chooses among those two lenses, a new pair is offered, and then another, and another until finally the lenses are so similar that one cannot choose between them. Unlike the situation at the eye doctor's, forced choice experiments generally ask the observer to choose among one or more options that are presented concurrently rather than sequentially, but the basic premise remains that one chooses among various alternatives the one that best fits the criteria by which one is judging. Sometimes, but not always, there is an option to judge the alternatives to be equal.

In medical imaging perception experiments, AFC experiments are often used when one wishes to determine which of two or more conditions are better for a particular purpose. Generally this approach is used in situations in which finding an abnormality is not the relevant task but rather determining under which tested condition the structure in question is better seen. Balassy et al. used this approach to study whether liquid crystal displays (LCD) or cathode ray tube (CRT) monitors were better for digital display of chest radiographs. Each image was displayed on both an LCD and a CRT. The monitors were side-by-side, so the observers could look back and forth from one to the other at will. The observers then graded the visibility of various anatomic landmarks as seen on one monitor versus the other. The observers were trained radiologists, so hopefully they knew where the landmarks were supposed to be and what they were supposed to look like, and the only question was whether the landmark was better seen on one monitor or the other (Balassy, 2005).

Balassy's experiment used a 2AFC design, and that seems to be the most common approach, but it is possible to have more alternatives. It is also possible, unlike in Balassy's experiment, to have an element of search involved. De Vries et al. performed an experiment to determine what level of tube charge would best portray colon polyps in CT colonography. They designed phantom colon rings that they scanned with different tube currents. Each ring had exactly one 6-mm polyp, and each image of the ring was divided into 8 segments. Observers were given the task of picking which of the segments contained the polyp. Thus, this was an 8AFC choice design in which observers were to seek out a phantom abnormality (De Vries, 2008).

In our experience, rank order experiments are not as common as AFC experiments. They are similar in that they ask the observer to decide which condition is best, but dissimilar in that they also ask which is second best, third best, etc. Good et al. studied the effect of data compression on mammography. Six digital versions of each image ranging from no compression to 101:1

compression were printed out on film, and radiologists were presented with film sets containing all six versions of a specific image and asked to rank them from best to worst and everything in between (Good, 1999).

Both AFC and rank order experiments absolutely require that the observer be able to remember one condition when viewing the other. The only memory-related requirement for the investigator is to make it as easy as possible for the observer to remember. If at all possible, arrange the images so they are in close physical proximity to one another and are displayed simultaneously, so the observer can look back and forth as many times as desired, with ease. If this is impossible, then a toggle approach can work, in which the observer can flip back and forth from one image to the other.

Besides making it easy to view the images back and forth in quick succession, investigators also want to make it easy for the observers to be sure which image is which, to minimize accidental mistakes in grading due to the observers becoming confused as to which image is A and which is B. One option that may work is to have the whole experiment conducted using a computer program in which the images are clearly labeled and the observers need to indicate which image they want to select as their response via the click of a mouse button. If time allows, the program can then show the image again to the observers and ask them to confirm their response. Even more simply, the observer can indicate the chosen image by clicking on the image itself. Another option is to have the images displayed side-by-side and clearly labeled so that when the score sheet asks something like, "do you prefer A or B," the observers can glance quickly back to the images to be sure of answering as intended. The answers may then be recorded on paper or in whatever way is convenient. As long as it is equally easy to identify each option and the target image is as likely to appear on one side as on the other, then mistakes due to confusion should not favor one option over another as the mistakes will hopefully cancel each other out; still, it is cleanest to avoid them if possible.

Sequential-viewing experiments are typically used when testing the effect of an add-on technology that in usual clinical practice would be an adjunct to another technology. Experiments testing computer-assisted detection (CAD) in mammography are often set up in this manner. Tchou et al. used sequential viewing in an experiment examining how much time it required to consult CAD while reading digital screening mammograms. and how often the use of CAD changed either the result of the interpretation or the radiologist's confidence in the interpretation. The radiologist interpreted the mammogram without CAD, committed to an interpretation and confidence level, and then immediately was shown the CAD image and allowed to make any change in interpretation or confidence level that seemed appropriate once both the mammogram and the CAD image were visible together (Tchou, 2010). There are numerous other articles in the literature describing sequential-viewing experiments, not all of which involve CAD or mammograms. Meisamy et al. used sequential viewing to assess how the addition of MR spectroscopy quantifying the concentration of total choline-containing compounds influenced radiologists' interpretation of breast MRIs. The radiologists viewed each MRI, interpreted it, and then were given the result of the MR spectroscopy and interpreted the study as a whole again (Meisamy, 2005).

In sequential-viewing experiments such as these, the observer normally grades an image alone and then gives another grading taking into account both the original image and the add-on, whether that be CAD or some other piece of information. When grading the two technologies together, it is vital that the observer be able to remember both at the same time and ideally be able to flip back and forth between them, just as would be done in clinical practice.

Experiments That Create Memory Bias

There are only a few types of observer-performance studies in which memory is not needed for the study to work, nor will it bias the results. Those are experiments using a non-human model observer, and those in which the observers do not see the same images more than once. These include designs in which a single imaging modality is compared to a non-imaging gold standard such as surgical findings. They also include those in which two or more imaging studies are compared but the different types are so removed from one another that there is really nothing to recognize on the second interpretation. For example, if ultrasound images are compared to MRI.

In essentially all other types of observer-performance experiments, memory for previously-viewed images can potentially bias the results. Memory can conceivably influence experimental results in different ways. The first and most obvious way is that an observer consciously recognizes an image, remembers his or her first reading, and more-or-less matches that reading on the second encounter. I (Tamara Haygood) became interested in memory as a factor in observer-performance experiments partly because of an experience I had acting as an observer in a colleague's experiment. The task was to hunt for pulmonary nodules on chest radiographs in two different viewing conditions. On the second viewing, I found two nodules on a radiograph and was about to hit the button that would mark that case as finished, when I suddenly realized that I had seen that radiograph before. Furthermore, I remembered that it had had three nodules (or at least I had thought it had three nodules),so I kept hunting until I found the third nodule, marked it, and then went on to the next case with a happy sense of triumph. So, in the case of that particular radiograph, my memory for it clearly influenced the results.

Conceivably, conscious recognition of an image could also affect results in the opposite direction. Observers could recognize an image and even remember the previous reading yet try so hard to avoid being biased by that memory as to talk themselves out of an interpretation that they might otherwise have made. There is also a possibility that observers can misremember a case. That is, they might believe erroneously that the image is the one they saw and interpreted before (i.e., a false memory) giving the remembered assessment rather than scrutinizing the image in front of them.

Memory could also influence results when recognition occurs at a less-than-conscious level. Kallergi conducted an experiment to determine the value of high-resolution head and neck PET/CT scans as an adjunct to whole-body PET/CT in evaluation of patients with thyroid cancer. She recruited two readers and had them interpret the whole-body PET/CTs. She had built into the experiment a time gap between that reading and the reading with the high-resolution CTs, but right after the first reading she went out of town. When she came back, she found that her readers had jumped in and within two days had done the second reading, expecting her to be pleased. She was not as pleased as they had hoped and told them she

suspected that the short time period between the two readings would influence the results due to memory for the first reading. They protested that they had not remembered the studies and there would be no bias from the short interval between readings.

At that moment, Dr. Kallergi conceived a new study. She persuaded the readers to wait a month and then re-read the PET/CT plus high-resolution PET/CT combination studies. She and her co-workers found that the results of the first two readings were similar, while there were significant differences between both of those readings and the third reading. Thus, even though the readers themselves had not consciously remembered their first readings and did not believe they influenced the second readings, they had nonetheless affected the reading done within two days (Kallergi, personal communication, 2017; Kallergi, 2012).

Ways to Mitigate Memory-related Bias

We will consider two methods to decrease the effect of memory-related bias on the outcome of an observer-performance experiment. One can either take steps to decrease memory for the images, or one can organize the experiment to cancel out the effects by using counter-balancing methods. Within each of these major categories, there are various steps that can be taken.

Counter-Balanced Methods - Ordering of Test Conditions

Many observer-performance studies compare performance in two different settings. They may differ either in the images themselves being different though still recognizable (pictures of chairs in black and white versus color photographs), or in something about the environment being different (pictures of chairs being viewed from up close or across the room). Suppose the observers are comparing black and white versus color images and the task is to decide whether the upholstery is tufted or not. Observers have to make it through the stack of chair pictures twice, once in black and white and once in color.

Investigators can present image set variations in three ways:

1. Every observer looks first at the same set of images and then at the other (all see the black and white images first and then the color images);
2. Every observer looks at one version and then the other, but which goes first and second is variable, with the number of times that each set is viewed first being equal; or
3. The black and white and color images are mixed together.

With any of these methods there is the possibility that an observer will remember an image, but the likelihood that memory might affect the results will vary with the design. In the first method, the color photos, always shown second, will be the only ones whose interpretation might be affected by memory of the first viewing. Therefore it is reasonably likely that the ordering of the image sets may affect the outcome. It is easy to imagine an observer looking at one of the color pictures, not immediately noticing the tufts but recognizing the beautiful Queen Anne chair portrayed, remembering that it was tufted and squinting a little bit harder to confirm – yes! There are tufts!

If the black and white photographs are followed immediately by the color photographs in the same session, another potential problem is that observers may change their reading methods and thresholds over the course of a long reading session. Taylor-Phillips and colleagues performed a second-look analysis of data from six observer-performance studies and found that in the four studies that included time information for individual readings, readers decreased the amount of time spent per case as they proceeded through the images. In the studies with reading sessions of 60 or 100 cases, there was also either a decrease in sensitivity for cases presented later in the session or an increase in specificity. This was not seen in the studies with 27 or 50 cases per session (Taylor-Phillips, 2015). Fatigue may also set in when there is a large number of cases, and that has been shown to adversely affect observer performance in reading cases (Reiner, 2012). Thus, if the observers in our hypothetical chair experiment are to look at a fairly large number of chair pictures, having all the black-and-white images shown before the color images can affect the results in other ways besides those directly related to memory. Therefore we would not recommend this method.

In the second ordering method the investigators have ensured that more-or-less half of the observers see the black and white photos first and the other half see the color images first. That will not prevent people from remembering the images, but it will prevent one set of images having an unfair advantage over the other based on this memory, as the advantages given one set with some readers versus the other set of images having an advantage with other readers will cancel each other out.

In the third ordering method, assuming the mixing is thorough enough, the memory effects for one image set versus the other should also cancel out. This assumes that the experiment is set up so that the observers cannot go back and change answers once they have committed.

Decreasing Recognition Memory - Ordering of Images within Sets

In our proposed tufted-chair experiment, once the investigators have decided how they are going to order the viewing of the black and white and color photographs, they also have to decide how to order the images within the sets. In this context, by a set of images, we mean a group of images that will be viewed sequentially in a single viewing session. In both the first method (everyone sees the same set first) and the second method (observers see different sets first), it will help to decrease recognition of images if the images in the two sets are in different orders. In memory, serial position effects are quite pronounced (Shiffrin, 1997). For example, observers are much more likely to recall or recognize items that they saw or heard first (primacy effect) and last (recency effect) in a series of items, than they are items in the middle. The primacy effect is less pronounced when the image set is large and the rate of presentation is fast. Recognition of one image can also prime an observer to recognize the next image, if the two are in the same order that they were in on the first viewing. This is something that parents and grade-school teachers all seem to know. Spelling lists may be given in one order, but they will be tested in another, and parents helping their children memorize the words will also vary the order. Being able to spell “that” when it comes right after “flat” is not the same thing as being able to spell it when it comes after “splat.” One would also want to mix up the images with one correct answer versus another. It would not do to have all the tufted chairs together and all the ones without tufts together. Any intelligent observer would catch on.

Mixing up the images can be done several different ways. The choice may depend partly on how many images one is dealing with, partly on how the images are displayed (electronic display or physical display – printed on actual film or paper), and partly on how the transition is handled from one image to the next and how the results are recorded.

Random mixing is probably the cleanest. Each image can be assigned a number, and there are free random-number generators available on line. (Try http://stattrek.com/statistics/random-number-generator.aspx, accessed 22 September 2017.) Images can then be shown in the order in which their numbers appear in the lists, and a new list is created for each new observer. This will work for any type of display. If, however, one is dealing with a relatively small number of images and a relatively small number of readings, one might make sure that the images appear in all possible permutations of that list. Be aware that some random-number generators are designed so that for a given number of cases they will always generate the same list. For example, if one asks for a random list of numbers between 1 and 50, the list may be "36, 18, 43, 9, 11 etc." Then if one asks a second time for a random list of numbers between 1 and 50, the list is again "36, 18, 43, 9, 11 etc." This will not work, as the idea, of course, is for the images to be in different order each time the observers see them.

Other methods produce what might be called pseudo-random mixing. Arthur de Smet and colleagues studied radiologists' ability to identify meniscal tears in MRIs (magnetic resonance imaging) of the knee. They compared MRI results with findings at arthroscopic surgery, so memory for images was not a factor, yet the study is interesting as an example of a reasonable method of pseudo-random mixing of cases. They included 200 cases shown on film and read by three observers. The cases were presented to the observers in order according to the patient's last name. That ordering scheme is clearly not truly random, but it should produce a reasonably mixed up assortment of normal MRIs and MRIs showing meniscal tears in one location or another, simply because a person's likelihood of having a meniscal tear has nothing to do with what the first few letters of their last name may be (De Smet, 1993). De Hoop et al. also sorted images by patient last name in an experiment comparing mammography to breast tomosynthesis (De Hoop, 2010). Since these methods will produce the same viewing order for each observer, they are best used in experiments where the images are viewed only once, or in which the types of images compared are quite different from one another, and therefore memory is not a factor.

If the experiment is run electronically, with images sorted and presented by computer, then it might be possible for the computer to take care of randomizing the images. In a study by Evans et al., the computer did the randomizing in a study of radiologists' ability to identify mammograms harboring breast cancer after a viewing time of only half a second. Every time a radiologist sat down to view the images, they were presented in a unique order so no two radiologists read them in the same sequence. The computer also kept track of the answers given by the observers, eliminating a potential source of confusion in record-keeping. (Evans and Haygood, 2016).

Decreasing Recognition Memory - Length of viewing

The longer a person has to stare at an image, the more likely it is that the image will be remembered. This is due both to the opportunity to actively rehearse the material, as well as the more passive aspect of the process of memory consolidation. Memory consolidation is a process by which a memory trace is stabilized either through synaptic (long-term potentiation that strengthens new synaptic connections formed during encoding) or system (process of reorganization of the neo-cortex connections based on hippocampal repeated activation) consolidation. Synaptic consolidation is more likely to be relevant to observer-performance studies than is system consolidation, as the latter applies more to repeated exposures to material occurring over days to weeks or months (Kandel, 2000). A longer viewing allows repeated visual fixation on whatever features the observer may find interesting, with each repetition contributing to placement of the image in long-term memory. It also allows the observer to look with some attention around the entire image and increases the likelihood that something will catch the observer's eye and will seem interesting enough that the image will be remembered.

Although decreasing the length of time the observers can view each image should also to a degree decrease memory, we would suggest that in general other reasons than memory should take precedence in deciding how long to allow observers to view each image. Our experiments investigating radiologists' ability to find mammographic abnormalities after a mere half second of viewing time were designed to understand the role of the instantaneous first impression or gist of an image in diagnosis, and therefore the half second viewing time served the main purpose of the experiments. Its only significance where memory was concerned was that it contributed to our confidence that we could show all the images in one sitting, in random order one right after the other, without the observers being able to recognize any that they were seeing for a second time (Evans and Haygood, 2016).

In experiments attempting to mimic normal clinical viewing circumstances, one would normally wish either to impose no time limit at all or to use a time limit that approximates the outer limits of the time normally taken for the task in the course of usual image interpretation. Kim et al, in a study of the performance of radiologists in detection of urinary stones on radiographs using film versus two digital display methods, did not limit viewing time. They wanted to simulate a normal viewing environment as closely as possible, and a time limit would have detracted from that intent. Differences in interpretation time with the three different display methods were also part of what they wanted to investigate, and a time limit would also have detracted from that goal (Kim, 2001). In this study, having no time limit served the main purpose of the investigation, and we presume that in the authors' opinions it was worth the small added risk that an observer might remember one of the images.

In a study aimed at determining the effect of observers' being or not being forewarned of a memory-related task, Haygood et al. used a viewing time limit of 40 seconds. This limit was chosen as this experiment involved interpretation of single-view chest radiographs, and the time limit was similar to limits that had been used by other authors in studies also looking at interpretation of single-view chest radiographs. It also was thought (and subsequently supported) to be adequate for interpretation. The only purpose the time limit served was to prevent the readers who had been forewarned of the memory task from prolonged staring with the intent of memorizing the image. As it turned out, that was not necessary as both the

forewarned readers and those who were not forewarned kept their interpretation time well within the limit (Haygood, 2013).

Decreasing Recognition Memory – Time Gaps

Although as stated previously, LTM is based on the formation of stable neural connections, either a memory itself or one's ability to access the memory clearly does fade with time. Think back to grammar school and try to remember the names of all the other students in your third-grade class. If you're like us, you can't. Dig out the class photo from that year. Now you have pictures of all the kids in your class, and you still probably can't name them all – but will likely remember more (the well-known recall vs recognition distinction). (No looking at the printed list of names – that's cheating.)

No doubt there are some readers smugly ticking off on their fingers the name of every single third-grade classmate. Most likely those readers are either quite young, so third grade was not terribly long ago, or they attended a very small school where the same cohort of students was together every year, so there was lots of practice with the names, so recalling third grade is the same as recalling twelfth grade.

For most of us, an attempt to remember our third-grade classmates is an adequate reminder of what we observe every day, that memories fade over time. The primary reasons that long-term memories fade are interference from new memories, competition between memories and a result of retrieval dynamics (e.g., the act of remembering inhibits at the same time the retrieval of related information) (Anderson, 1994; Squire, 1989). Studies that have tracked the forgetting rate in long-term memory observe that forgetting is not uniform, with most of it happening within the first month after formation of a memory and then levelling off thereafter (Landauer, 1986). For investigators planning an observer-performance study, the point of incorporating a time gap between readings is to give an opportunity for any long-term memories that may have formed of the images to diminish.

So what is an adequate time gap?

I (Tamara Haygood) am a practicing radiologist. At one time I was using a dictation system that allowed doctors to occasionally choose, out of a queue of studies, one that had already been interpreted. An interpreted study would drop out of the queue within a second or so of the time the dictation was completed, but in that second, it was possible for the same radiologist who had just interpreted it to open it again and start another dictation (i.e., on the same image!). I actually did that several times. I would then get a phone call from an administrator pointing out that there were two dictations on Mrs. Jones – which one did I really want to use? This usually happened with chest radiographs that had no striking findings and were essentially normal, so there was nothing that caught my eye and formed a trigger for memory. When I mentioned this to another radiologist in my practice, he replied, "Oh, we all do that." If all cases were like these, no time gap would be needed at all.

Another time, I was interpreting a set of forearm radiographs. I hunted through the patient's images on the PACS (Picture Archiving and Communications System) and found the original examination from four years earlier. I recognized that image immediately and remembered the patient's name and the type of tumor that the patient had had. How could I, who was capable of forgetting and re-dictating a radiograph that I had just interpreted, also remember in some detail another radiograph after four years? Simple. This patient had a metastasis from renal cell carcinoma. This metastasis had demanded so much blood supply that the nutrient canal (the pathway through the bone that allows the artery to enter and supply blood to the interior of the bone) had noticeably enlarged. Fascinating, at least to me. If all cases were like this one, essentially no amount of time gap would prevent conscious memory of images. (Figure 2.)

Figure 2. 67-year-old man with renal cell carcinoma. A. AP radiograph of the proximal right forearm shows a lytic lesion caused by a metastasis. The nutrient canal (arrow) measures 2.7mm in diameter. B AP radiograph of the normal, proximal left forearm in the same gentleman. The nutrient canal measures 1.7mm in diameter.

The truth, of course, is somewhere in between. Charles Metz, in an article that provided practical suggestions on how to design and run an observer-performance study, said that readings by a single observer of the same image should be separated by as much time as possible (Metz, 1989). Since then several published studies have shed additional light on the subject by investigating radiologists' memory for radiographs they have encountered.

Hardesty et al. questioned whether it was reasonable to include in observer performance studies images that had originally been interpreted by the same people who would serve as observers. They gathered 33 mammograms and 4 radiologists as readers. Among the 33 mammograms were 5 from each reader showing a cancer that he or she had correctly identified and 2 showing a cancer that the reader had not identified on original interpretation. This gave a total of 28. 5 Five mammograms that had been interpreted by a radiologist who did not serve as a reader. These 5 were called back due to an abnormality eventually proven to be benign. The investigators mixed the mammograms together and then asked the 4 radiologists to interpret them. The mammograms had originally been interpreted two to three years before the experiment was run. As each case was shown, the observers were asked not only to interpret it but also to indicate if it were one that they had originally interpreted. Only one radiologist correctly identified exactly one case that he had reported, and he gave a different opinion on it in the experimental setting than he had at the time he reported on it for clinical purposes. Hardesty et al. concluded that investigators may reasonably incorporate into an observer-performance study patient images originally interpreted by the same people serving as observers, without concern that the

observers will remember the images or that their encountering studies they had previously interpreted would bias the results (Hardesty, 2005). We agree with this conclusion. As the story of the renal-cell metastasis illustrates, radiologists can remember an image for a long period of time, but we believe that is a sufficiently rare event that it is not a practical concern that a radiologist observer will recognize something interpreted years earlier.

Hardesty's article applies to a time gap between original interpretation and when radiologists would encounter an examination in an experimental setting. It also implies that a time gap within an experiment of two to three years would be sufficient. Although Fuhrman et al. included a two-year time gap between viewings in an experiment involving rib fracture detection on chest radiographs, such a long time gap is not practical in many situations (Fuhrman, 2002). When observations are made in settings such a scientific meeting, all viewings have to be completed within a short period, usually a week or less. What is the likely effect of memory in experiments with shorter time courses?

In Ryan et al.'s experiment, radiologists were asked to distinguish new from old chest radiographs. The two viewings took place one to three days apart. When data from all 24 participating radiologists were pooled, radiologists were correct in their classification of images as new or old 67% of the time, with ability to make the distinction varying among individuals (Ryan, 2011).

Hillard et al. showed slides of 20 chest radiographs to radiologists with varied levels of experience, including 5 first-year residents, 4 junior staff radiologists with an average of 3.5 years of experience, and 6 senior staff radiologists with an average of 18.5 years of experience. Each slide was shown for 500 msec, and immediately afterwards the radiologists were shown 40 slides of chest radiographs, half of which were new and half of which were those previously shown. The radiologists were asked to categorize the images as new or old. Correct categorization ranged from 45% to 71% (Hillard, 1985).

Evans et al., tested visual recognition memory of 12 board certified radiologists with 108 anonymized chest radiographs and found that their memory for them, when tested right after having an opportunity to memorize them, was poor but significantly above chance. They correctly recognized 65% of images as having been seen before. When they used a wider variety of images (108 anonymized musculoskeletal radiographs of varied body parts) their performance improved marginally with 72% correctly recognized. This improvement completely disappeared with a gap between study and test phase of an average of 50 days. The radiologists' performance then dropped to chance (50% recognition rate) (Evans and Marom, 2016).

These studies suggest that observer-performance experiments utilizing a moderate number of images (up to 40) can expect a low level of conscious recognition of images when a second viewing occurs either immediately after the first viewing or within three days. A time gap of seven weeks should eliminate recognition of images. Indeed, Landauer's work suggests that a time gap of a month should be sufficient for most of this memory decay to occur (Landauer, 1986). The studies by Evans, Hillard, and Ryan were all performed using one image per case. It is not known how sets of multiple images shown together would affect memory. For example, if one showed simultaneously the AP (anteroposterior), lateral, and oblique radiographs of the

ankle, would the additional images change the likelihood that observers would remember the set as a whole compared with showing only one of the views?

There is also a possibility that even though there is no conscious explicit recognition of images there is implicit memory (i.e., facilitation of test performance without conscious recollection) for them. Studies in different contexts have shown that observers show priming effects (e.g., greater facility) with material they have learned before but cannot explicitly recall or recognize (Lewandowsky, 2014). In everyday life, implicit memory is most evident in so called procedural memories like tying one's shoes or riding a bike (Schacter, 1993). Therefore it is possible that readers of medical images might demonstrate faster reading times or ease of processing with images they have encountered before without being conscious of ever seeing them. How much is this present and how much of an important issue it is for studies in medical image perception has not been studied and is still unknown.

The studies discussed above were concerned with observers' ability, on viewing a medical image, to determine if it was one they had previously encountered. Conscious recognition of an image, however, is not terribly relevant if recognition does not affect the interpretation. One would expect that recognition would be accompanied by an increase in consistency of interpretation between readings for recognized images as compared with consistency of interpretation for unrecognized images. A few investigators have looked at interpretation consistency and have had mixed results. In Hardesty's experiment the one mammogram that was recognized received a different interpretation on its second viewing (Hardesty, 2005). Kallergi did not test for conscious recognition, but did find greater consistency of interpretation between viewings just a day or two apart and viewings separated by several weeks (Kallergi, 2012). Ryan et al. inquired about conscious recognition of chest radiographs containing central lines. Comparison of first and second interpretations of central line position for repeated images that were recognized versus those that were not recognized did not demonstrate any increase in consistency of interpretation for recognized images. Indeed there was a non-significant trend for more consistency of interpretation for unrecognized images (Haygood, 2012). As these studies are somewhat contradictory, we would say the jury is still out as to the likely practical effect of a participant's recognition of a previously-viewed image.

Decreasing Recognition Memory – Elimination of Extraneous Information

Memories can be triggered by many different things. Smells are especially inclined to bring up memories of past events, but sights can also. I (Tamara Haygood) have a small collection of things that were once my parents' sitting on a mantel in my home. A pair of framed bird paintings, a level that my father used and that had once been his father's, a green glass vase that once sat in my mother's living room. A quick glance at any of these objects is enough to call up childhood memories.

In observer-performance studies, anything on the image can trigger a memory from a previous viewing session. One way to mitigate this effect is to exclude unnecessary information. When the experiment uses images obtained from real human beings, the first information normally excluded is identifying information about the individual people whose images are being used. Often this must be excluded anyway for the sake of the individual's privacy. Thus, name,

address, phone number, etc. are not revealed to the observers. Sometimes the nature of the experiment requires revealing some limited information such as age, gender, and the reason the images were obtained.

One should also be mindful of extraneous information included in the images themselves. Aa my story of the enlarged nutrient canal suggests, an eye-catching normal variant or an abnormality other than the precise type of abnormality relevant to the particular experiment can trigger memory. In memory research this kind of information or image characteristics are referred as *conceptual hooks* (Brady, 2011). When material to be remembered can be linked to existing knowledge or can be semantically labeled (e.g., a mammogram with architectural distortion) it significantly improves our memory for the same. Hollingworth and Henderson showed that observers encode and remember for a longer time material with very distinct details or those with elements that are inconsistent with the overall context of the image (Hollingworth, 2003). The issue of extraneous, inconsistent and distinct details in medical images is further complicated by the fact that in many observer-performance studies, these images are being viewed by experts who, because of their expertise, may easily find conceptual hooks in an image that would lack conceptual hooks for a novice observer.

One's ability to avoid showing eye-catching normal variants or abnormalities can be constrained by the ease with which one can find images showing the abnormality or anatomy being tested. In Ryan et al.'s experiment, in which observers were asked to determine whether a central venous access line resided in the superior vena cava or the azygos vein, a chest radiograph was included that depicted a patient with a well-known yet fairly unusual normal variant in which the azygos vein is suspended inside an envelope of pleura that indents the right upper lobe, creating a so-called azygos fissure and azygos lobe. The central line curled right into the vein where the vein is suspended at the bottom of the fissure. This is a sufficiently eye-catching appearance that a radiologist is likely to remember it after seeing it. This would be an image to avoid including if possible. It was included mostly because azygos placements are not common, and therefore there were only a limited number to choose among. In the case of that experiment, only a subset of the images were shown twice, and this image was not among them. It was shown to each observer only once (Ryan, 2011).

Figure 3. Chest radiograph showing a central venous access line entering the azygos vein. A. Full image. B. Image coned to the area of interest. The azygos fissure is marked by arrowheads, and an arrow points to the line entering the azygos vein.

Although variation in the type of image and its anatomy can influence the ability of radiologists to recognize repeated images, it is not clear how much influence the presence or absence of

abnormalities and type of abnormality present on the images may have. In Ryan's study, chest radiographs with a larger number of abnormalities and those with abnormalities that are clinically more significant were more often correctly classified as new or old than those that were either normal or contained a relatively common abnormality or one unlikely to indicate severe disease (Ryan, 2011). For example, a chest radiograph with a lung mass that could indicate lung cancer would be correctly classified more often than a chest radiograph with a mild spinal malalignment (scoliosis), which the radiologists would have realized is a common finding and not life-threatening. In Hillard et al.'s study, the more experienced faculty radiologists remembered best the abnormal images (71% correct classification), while they did not do well at distinguishing new from old normal radiographs (48% correct classification.) The first-year residents remembered best the normal radiographs (67% correct classification), while they did not do well at distinguishing new from old abnormal radiographs (55% correct classification.) The less experienced staff radiologists' performance with normal radiographs was similar to that of the more experienced staff (45% correct classification), and their performance with abnormal radiographs was similar to that of the residents (57% correct classification.) Hillard et al. believed this was related to development over time of schemata, or a mental image of chest radiographs and the range of what would be considered normal (Hillard, 1985).

Ryan and Hillard, therefore, both found that, at least among experienced radiologists, old and new images can be distinguished from one another better when they contain an abnormality. In our study in 2016, however, we found that radiologists' ability to distinguish new from old chest radiographs did not correlate with the presence or absence of an anatomic abnormality (Evans and Marom, 2016). Only when the material to be remembered was more varied did the fact that an image contain an abnormality result in better memory for that image.

If one wishes to be precise, all of the chest radiographs shown by Ryan et al. were abnormal in that they all contained a central venous access catheter, and these catheters are not really "normal" at all. The relevant finding that the radiologists was to look for was misplacement of the catheter into the azygos vein instead of the superior vena cava. That was present in half of the images. Azygos placement of a central line is fairly uncommon, so when it is found in clinical practice, it is rather eye-catching. In this experiment, however, azygos malposition of the catheter was not associated with improved ability of the radiologists to determine if the image was new or old (Ryan, 2011). This is probably because in the specific setting of this experiment, azygos placements, though uncommon clinically, were quite common, being present in half the images. This suggests that the presence or absence of whatever abnormality is being searched for may not, itself, provide a trigger for memory. This may, however, not be true in less enriched image sets. For example, if the azygos placements had only been present in 5% or 10% of cases, it would not be surprising if they then attracted more attention and were remembered. This possibility has not, however, been tested to our knowledge.

In the Ryan study (Ryan, 2011) the correlation between the number and severity of abnormalities and the radiologists' ability to distinguish previously-viewed from new images suggested that it might be possible to make reasonably accurate guesses in advance as to how easy an image might or might not be to recognize. Another study suggests that it might be a bit more complicated. A collection of 108 frontal chest radiographs was gathered. A random assortment of 72 were shown to each of 12 board-certified radiologists. The images were each visible for 3

seconds. The radiologists knew they would be tested on their memory for the images. Immediately afterwards each radiologist was shown another set of 72 images, half of which were randomly selected from the original set of 72 and half of which were new. They were asked to identify the new versus previously-viewed images (Evans and Marom, 2016).

The same 108 chest radiographs were also given as slides to three different chest radiologists who did not participate as observers in the study. These three radiologists were asked to sort the images into three different sets, one third of the images in each set. Each radiologist sorted the images independently. The sets were to contain those images that each radiologist thought would be easy to recognize, difficult to recognize, or of intermediate difficulty. There was no correlation between the ability of participants to distinguish previously-viewed from new images and the predicted memorability of the images, as judged by the other three (Evans and Marom, 2016).

In a closely related experiment, 108 varied musculoskeletal radiographs (spines, knees, arms, etc.) were shown to radiologists in a procedure similar to that described above. These were also sorted into groups thought to be easy to recognize, difficult to recognize, or of intermediate difficulty. With the musculoskeletal radiographs, there was a positive correlation between the opinion of the three radiologists who estimated the likelihood that a radiograph would be remembered and performance of the viewing radiologists. In other words, radiographs that were considered to be difficult to recognize proved actually to be difficult, and those that were expected to be easy actually were easy. There was also a positive correlation between estimated ease of recognition and the presence of an abnormality on the radiographs. The images with abnormalities were thought to be (and actually were) easier to recognize than those that were normal (Evans and Marom, 2016).

Conclusion

What recommendations can we make to investigators planning an observer-performance study? What steps might they reasonably take to decrease the chance that observers' memory for the images being shown might deleteriously alter the results of the experiment?

If the experiment will use AFC, rank-order, or sequential-viewing methods, then make it as easy as possible for the observers to remember the first-viewed images when assessing later images. For AFC and rank-order methods, display the images simultaneously so observers can look back and forth from one to another at will. For sequential-viewing experiments, do the same with the add-on and original images once the observer has committed to an interpretation of the original images. Adopt a scoring method that will minimize any confusion as to which image is which.

If the experiment will use the more common methodology in which identical or relatively similar images are viewed in two or more different conditions, then measures should be taken to decrease the likelihood that the observers' memory for the images may affect the outcome. The simplest means to this end is to counter-balance image presentation so that any advantage that memory may confer on one tested condition with one observer will be cancelled out by a similar memory advantage conferred on the other tested condition with another observer. Counter-

balancing can be done both at the level of the tested conditions and at the level of the order in which individual images are shown.

When using medical images taken from actual patients, exclude images that have unusual incidental normal variants or abnormalities that are either unusual or an indicator of serious illness. This is especially important if the observers are to be medical experts as such incidental findings will attract their attention and serve as a conceptual hook that will promote memory for the image. One's ability to exclude these eye-catching abnormalities will depend partly on how many images one has to choose from. If one is testing radiologists' ability to diagnose a fairly rare lesion, one may not be able to be especially picky about what images to use that show the relevant abnormality.

Include as many images in your test set as is reasonably feasible, taking into consideration the time your observers have available for helping you and the ease with which appropriate images can be found. The larger the sample size, the less likely it is that observers will remember specific images. Consider limiting viewing time, if doing so will serve the other purposes of your investigation.

A time gap between viewings can be helpful, but since conscious memory of images in relatively homogenous image sets (all images are frontal chest radiographs, for example) is fairly poor even with no time gap, we consider a time gap to be less important than the other recommendations made above. In particular, counter-balanced methodology to negate any effect of memory is most helpful and can allow you to dispense with a time gap if the circumstances under which the experiment is to be run do not allow for much time between viewings. If you do use a time gap, anything more than seven weeks is probably superfluous, and three or four weeks is most likely adequate.

References

1. Allard, F., Graham, S., Paarsalu, M. E., et al. (1980). Perception in sport: Basketball. *J Sport Psychol,* 2(1), 14-21.

2. Alloway, T. P., Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *J Exp Child Psychol*, 106, 20–29.

3. Anderson, M. C., (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *J Exp Psychol Learn,* **20**(5): 1063.

4. Baddeley, A. D. (1986). Working memory. Oxford, UK: Clarendon Press.

5. Baddeley, A. D., (2007). Working memory, thought, and action (Vol. 45). OUP Oxford.

6. Baddeley, A. D., Hitch, G. J. (2000). Development of working memory: Should the Pascual-Leone and the Baddeley and Hitch models be merged? *J Exp Child Psychol*, 77(2), 128-137.

7. Baddeley, A. D., Hitch, G. J. (2017). Is the Levels of Processing effect language-limited? *J Mem Lang*, *92*, 1-13.

8. Balassy, C., (2005). Flat-panel display (LCD) versus high-resolution gray-scale display (CRT) for chest radiography: an observer preference study. *Am J Roentgenol* **184**(3): 752-756.

9. Brady, T. F., Konkle, T., Alverez, G. A., et al. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *J Vision* **11**(5): 4, 1-34.

10. Brady, T. F., Konkle, T., Alvarez, G. A., et al. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci U S A* 105, 14325–14329.

11. Chase, W. G., Ericsson, K. A. (1982). Skill and working memory. *Psychol Learn Motiv*, 16, 1-58.

12. Chase, W. G., Simon, H. A. (1973). Perception in chess. *Cognitive Psychol,* 4(1), 55-81.

13. Cohen, M. A., Evans, K. K., Horowitz, T. S., et al. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonom B Rev*, 18(3), 586-591.

14. Daneman, M., Carpenter, P. A. (1980). Individual differences in working memory and reading. *J Verb Learn Verb Be*, 19, 450–466.

15. de Hoop, B., De Boo D.W., Gieterna H.A., et al. (2010). Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. *Radiology* **257**(2): 532-540.

16. de Smet, A.A., Norris M.A., Yandow D.R., et al. (1993). Diagnosis of meniscal tears of the knee with MR imaging: effect of observer variation and sample size on sensitivity and specificity. *Am J Roentgenol,* **160**: 555-559.

17. de Vries, A.H., Venema, H.W., Florie, J., et al. (2008). Influence of tagged fecal material on detectability of colorectal polyps at CT: phantom study. *Am J Roentgenol*, 2008, W181-W189.

18. Evans, K. K., Cohen, M. A., Tambouret, R., et al. (2011). Does visual expertise improve visual recognition memory? *Atten Percept Psycho,* 73(1), 30-35.

19. Evans, K. K., Haygood T.M., Cooper J., et al. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proc Natl Acad Sci U S A,* **113**(37): 10292-10297.

20. Evans, K. K., Marom E.M., Godoy M.C.B., et al. (2016). Radiologists remember mountains better than radiographs, or do they? *Journal of Medical Imaging,* **3**(1): 011005.

21. Frey, P. W., P. Adesman (1976). Recall memory for visually presented chess positions. *Mem Cognition,* **4**(5): 541-547.

22. Fuhrman, C. R., Britton C.A., Bender T.et al. (2002). Observer performance studies: detection of single versus multiple abnormalities of the chest. *Am J Roentgenol* **179**: 1551-1553.

23. Fukuda, K., Vogel, E. K., Mayr, U., et al. (2010). Quantity not quality: The relationship between fluid intelligence and working memory capacity. *Psychonom B Rev*, 17, 673–679.

24. Gardiner, J. M. (2001). Episodic memory and autonoetic consciousness: a first–person approach. *Philos T R Soc B,* **356**(1413): 1351-1361.

25. Good, W.F., Sumkin, J.H., Dash, N., et al (1999). Observer sensitivity to small differences: a multipoint rank-order experiment. *Am J Roentgenol*, 175:275-278.

26. Hardesty, L. A., Ganott M.A., Hakim C.M., et al. (2005). "Memory effect" in observer performance studies of mammograms. *Acad Radiol* **12**(3): 286-290.

27. Haygood, T.M., Liu, M.A.Q., Galvan, E.M., et al. (2013). Memory for previously viewed radiographs and the effect of prior knowledge of memory task. *Acad Radiol*, 10:1598-1603.

28. Haygood, T.M., Ryan, J., Liu, M.A.Q., et al (2012). Image recognition and consistency of response. *Proc SPIE 8318, Medical Imaging 2012: Image perception, observer performance, and technology assessment*.

29. Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends Cogn Sci*, *7*(11), 498-504**.**

30. Hillard, A., Myles-Worsley, M., Johnston W, et al. (1985). The development of radiologic schemata through training and experience: A Preliminary Communication. *Invest radiol* **18**: 422-425.

31. Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *J Exp Psychol Learn*, 31, 396–411.

32. Hollingworth, A., Henderson, J. M. (2003). Testing a conceptual locus for the inconsistent object change detection advantage in real-world scenes. *Mem Cognition*, 31, 930–940.

33. Isola, P., Xiao, J., Parikh, D., et al (2014). What makes a photograph memorable? *IEEE T Pattern Anal*, 36(7), 1469-1482.

34. Jolicoeur, P. (1985). The time to name disoriented natural objects. *Mem Cognition*, 13(4), 289-303.

35. Kane, M. J., Bleckley, M. K., Conway, A. R. A., et al. (2001). A controlled-attention view of working-memory capacity. *J Exp Psychol Gen* 130, 169–183.

36. Kallergi, M., Pianou, N., Georgakopoulos, A., et al (2012). Quantitative evaluation of the memory bias effect in ROC studies with PET/CT. *Proc SPIE* (Vol. 8318, pp. 83180D1-8).

37. Kandel, E. R., Schwartz, J.H., Jessell, T.M. et al. (2000). Principles of neural science, McGraw-Hill New York (Vol. 4, pp. 1227-1246).

38. Kim, A. Y., Cho K.S., Song K., et al. (2001). Urinary calculi on computed radiography: comparison of observer performance with hard-copy versus soft-copy images on different viewer systems. *Am J Roentgenol* **177**(2): 331-335.

39. Konkle, T., Brady, T. F., Alvarez, G. A., et al. (2010b). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychol Sci*, 21, 1551–1556.

40. Koutstaal, W. (2003). Older adults encode—but do not always use—perceptual details: Intentional versus unintentional effects of detail on memory judgments. *Psychol Sci*, 14(2), 189-193.

41. Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Sci* **10**(4): 477-493.

42. Lewandowsky, S. (2014). Implicit memory: Theoretical issues, Psychology Press.

43. Meisamy, S., Bolan P.J., Baker, E.H., et al. (2005). Adding in vivo quantitative ^{1}H MR spectroscopy to improve diagnostic accuracy of breast MR imaging: preliminary results of observer performance study at 4.0 T. *Radiology* **236**: 465-47.

44. Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annu Rev Psychol,* 32(1), 89-115.

45. Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* **24**: 234-245.

46. Reiner, B. I. and E. Krupinski (2012). The insidious problem of fatigue in medical imaging practice. *J Digit Imaging* **25**(1): 3-6.

47. Ryan, J. T., Haygood T.M., Yamal, J., et al. (2011). The "memory effect" for repeated radiologic observations. *Am J Roentgenol* **197**: W985-W991.

48. Schacter, D. L., (1993). Implicit memory: A selective review. *Annu Rev Neurosci* **16**(1): 159-182.

49. Schacter, D. L., Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter and E. Tulving (Eds.), Memory systems. Cambridge, MA: MIT Press.

50. Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *J Verb Learn Verb B*, 6, 156–163.

51. Shiffrin, R. M. and M. Steyvers (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonom B Rev*, **4**(2): 145-166.

52. Sperling, G. (1963). A model for visual memory tasks. *Hum Factors*, 5(1), 19-31.

53. Squire, L. R. (1989). On the course of forgetting in very long-term memory. *J Exp Psychol Learn* 15(2): 241.

54. Standing, L. (1973). Learning 10,000 pictures. *Q J Exp Psychol*, 25, 207–222.

55. Taylor-Phillips, S., Elze M.D., Krupinski, E. A., et al. (2014). Retrospective review of the drop in observer detection performance over time in lesion-enriched experimental studies. *J Digit Imaging* DOI 10.1007/s10278-014-9717-9.

56. Tchou, P.M., Haygood, T.M., Atkinson, E.N., et al. (2010). Interpretation time of computer-aided detection at screening mammography. *Radiology*, 257:40-46.

57. Tulving, E. (2000). Concepts of memory. The Oxford handbook of memory, 33-43. Brady, T. F., et al. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *J Vision* **11**(5): 4-4.

58. Tulving, E. (2002). Episodic memory: from mind to brain. *Annu Rev Psychol*, 53(1), 1-25.

59. Vogt, S., Magnussen, S. (2007). Expertise in pictorial perception: eye-movement patterns and visual memory in artists and laymen. *Perception*, *36*(1), 91-100.

60. Voss, J. F., Vesonder, G. T., Spilich, G. J., et al. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *J Verb Learn Verb Be*, 19(6), 651-667.

61. Wiseman, S., Neisser, U. (1974). Perceptual organization as a determinant of visual recognition memory. *Am J Psychol*, 87:675-681.

Figure Legends

Figure 1. A is a photograph of author Karla Evans. It is easily recognizable as a picture of a human face. B is a stylized, abstract rendering of the photograph of Dr. Evans. C is an abstract figure not based on a human face. Although B is clearly not an actual photograph of a human being, it bears enough resemblance to a face that we believe most people would recognize it as human-like and would have an easier time remembering and recognizing it than C.

Figure 2. 67-year-old man with renal cell carcinoma. A. AP radiograph of the proximal right forearm shows a lytic lesion caused by a metastasis. The nutrient canal (arrow) measures 2.7mm in diameter. B AP radiograph of the normal, proximal left forearm in the same gentleman. The nutrient canal measures 1.7mm in diameter.

Figure 3. Chest radiograph showing a central venous access line entering the azygos vein. A. Full image. B. Image coned to the area of interest. The azygos fissure is marked by arrowheads, and an arrow points to the line entering the azygos vein.